Title: Tracking SARS-CoV-2 Spike Protein Mutations in the United States (2020/01 – 2021/03) Using a Statistical Learning Strategy

Authors: Lue Ping Zhao[1]*, Terry P. Lybrand[2,3], Peter B. Gilbert[4], Thomas R. Hawn[5,6], Joshua T. Schiffer[4,5], Leonidas Stamatatos[4,6], Thomas H. Payne[5], Lindsay N. Carpp[4], Daniel E. Geraghty[7] and Keith R. Jerome[4]

Affiliations:

[1] Public Health Sciences Division, Fred Hutchinson Cancer Research Center; Seattle, WA, USA.

[2] Quintepa Computing LLC; Nashville, TN, USA.

[3] Department of Chemistry; Department of Pharmacology, Vanderbilt University; Nashville, TN, USA.

[4] Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center; Seattle, WA, USA.

[5] Department of Medicine, University of Washington School of Medicine; Seattle, WA, USA.

[6] Department of Global Health, University of Washington; Seattle, WA, USA.

[7] Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle; WA, USA.

*Corresponding author. Email: lzhao@fredhutch.org.

Abstract: The emergence and establishment of SARS-CoV-2 variants of interest (VOI) and variants of concern (VOC) highlight the importance of genomic surveillance. We propose a statistical learning strategy (SLS) for identifying and spatiotemporally tracking potentially relevant Spike protein mutations. We analyzed 167,893 Spike protein sequences from US COVID-19 cases (excluding 21,391 sequences from VOI/VOC strains) deposited at GISAID from January 19, 2020 to March 15, 2021. Alignment against the reference Spike protein sequence led to the identification of viral residue variants (VRVs), i.e., residues harboring a substitution compared to the reference strain. Next, generalized additive models were applied to model VRV temporal dynamics, to identify VRVs with significant and substantial dynamics (false discovery rate q-value <0.01; maximum VRV proportion > 10% on at least one day). Unsupervised learning was then applied to hierarchically organize VRVs by spatiotemporal patterns and identify VRV-haplotypes. Finally, homology modelling was performed to gain insight into potential impact of VRVs on Spike protein structure. We identified 90 VRVs, 71 of which have not previously been observed in a VOI/VOC, and 35 of which have emerged recently and are durably present. Our analysis identifies 17 VRVs ~91 days earlier than their first corresponding VOI/VOC publication. Unsupervised learning revealed eight VRV-haplotypes of 4 VRVs or more, suggesting two emerging strains (B1.1.222 and B.1.234). Structural modeling supported potential functional impact of the D1118H and L452R mutations. The SLS approach equally monitors all Spike residues over time, independently of existing phylogenic classifications, and is complementary to existing genomic surveillance methods.

41 **Main Text:**

42 **INTRODUCTION**

43 Severe Acute Respiratory Syndrome coronavirus 2 (SARS-CoV-2), the pathogen
44 responsible for the global Covid-19 pandemic, is an RNA virus and thus prone to replication
45 errors (*1*). Replication errors that yield nonsynonymous amino acid (AA) substitutions, or
46 nucleotide insertions or deletions that cause a frame shift and alter the subsequent coding
47 sequence, can lead to a variety of outcomes. If the resulting mutations have detrimental effects
48 on fitness, or if they have neutral effects on fitness and undergo stochastic extinction, variants
49 harboring these mutations fail to become established in the population. However, mutations that
50 confer a fitness advantage can rapidly become dominant in a population. For SARS-CoV-2,
51 there are three classes of variant: Variant of Interest (VOI), Variant of Concern (VOC), and
52 Variant of High Consequence (VOHC). The CDC is currently monitoring and characterizing 8
53 VOIs (B.1.526, B.1.526.1, B.1.525, P.2, B.1.617, B.1.617.1, B.1.617.2, B.1.617.3) and 5 VOCs
54 (B.1.1.7, P.1, B.1.351, B.1.427, B.1.429) in the United States (*2*). VOCs show specific attributes
55 such as increased transmissibility (*3-7*), increased resistance to neutralization by antibodies
56 elicited through natural infection (*3, 8-10*), and/or increased resistance to neutralization by
57 vaccine-elicited antibodies (*10-12*), and have already influenced vaccine development, evidenced
58 by the current planning of clinical trials to test variant-adapted vaccines (*13*). While no VOHCs
59 have yet been identified, it remains possible that such variants – i.e. variants that can effectively
60 evade natural or vaccine-induced immunity – may yet emerge (*14, 15*). The identification of
61 VOHCs could necessitate the introduction of more stringent public health guidelines and/or spur
62 further treatment and vaccine development.

Genomic surveillance is critical for tracking the emergence and spread of new variants. Such surveillance can be accomplished via a variety of approaches, such as phylogenic analysis (*3, 16*). In this approach, new viral sequences are classified to existing lineages identified by PANGO (*17*), subsets of samples with the same branches are identified, and variant frequencies are counted to identify new variants. The NextStrain methodology (*18*) can model dynamic changes of variant proportions, while an alternative approach aligns sequence data to a matrix of binary indicators for the presence of variants, and systematically evaluates each mutant as a potential variant (*19*). Leveraging the analytic approach of single nucleotide polymorphisms (SNPs), variants have been identified by assessing linkage-disequilibrium (*20*) or similar SNP-based identification and analysis (*21*). However, with the exception of the NextStrain methodology (*3*), these methods do not directly take into account sequence collection time, nor explicitly incorporate highly granular geographic information. Moreover, these methods take a holistic view of the viral genome. Thus, there is a need for complementary approaches for detecting and characterizing Spike mutations of potential public health importance that may be missed, or detected later, by existing genomic surveillance methods.

To meet this need, we describe a statistical learning strategy (SLS) using generalized additive models, unsupervised learning techniques, and single nucleotide polymorphism (SNP) methodologies for identifying and spatiotemporally characterizing viral residue variants (VRVs), a term we use to describe AA positions in the Spike protein where a mutation is significantly present in a given geographic area. The SLS method generates pertinent statistics for reproducible scientific inference and facilitates visual representation of results for intuitive interpretation. Using publicly available SARS-CoV-2 sequences from US COVID-19 cases that were not assigned to a VOI or VOC lineage, we apply our method to identify and

spatiotemporally characterize, within individual US states/territories, VRVs in the Spike protein. We also apply standard homology modeling methods to highlight individual AA mutations with the potential to impact Spike protein structure and/or function.

RESULTS

Ab Initio Discovery of VRVs

We first applied the SLS method to identify VRVs separately in each state/territory (Fig. S1). The decision to compartmentalize VRV discovery by state/territory was partially based on the fact that domestic travel restrictions have varied over the course of the pandemic, with nearly half of all states having imposed some type of interstate travel restriction (*22*), leading to the hypothesis that VRVs may follow state/territory-specific temporal dynamics. The identified VRVs showed a range of dynamic patterns across the different states/territories (Fig. S2), exemplified by the five different trajectories taken by the V382, L452, T478, P681, and T732 VRVs in California (Fig. 1A). The relative abundance of V382 started rising on day 250, exceeded 10% on day 259, and fell below 10% on day 275. L452 emerged on day 310, exceeded 10% on day 390, and exhibited a positive trajectory thereafter. Three other VRVs (T478, P681, T732) had similar trajectories to L452.

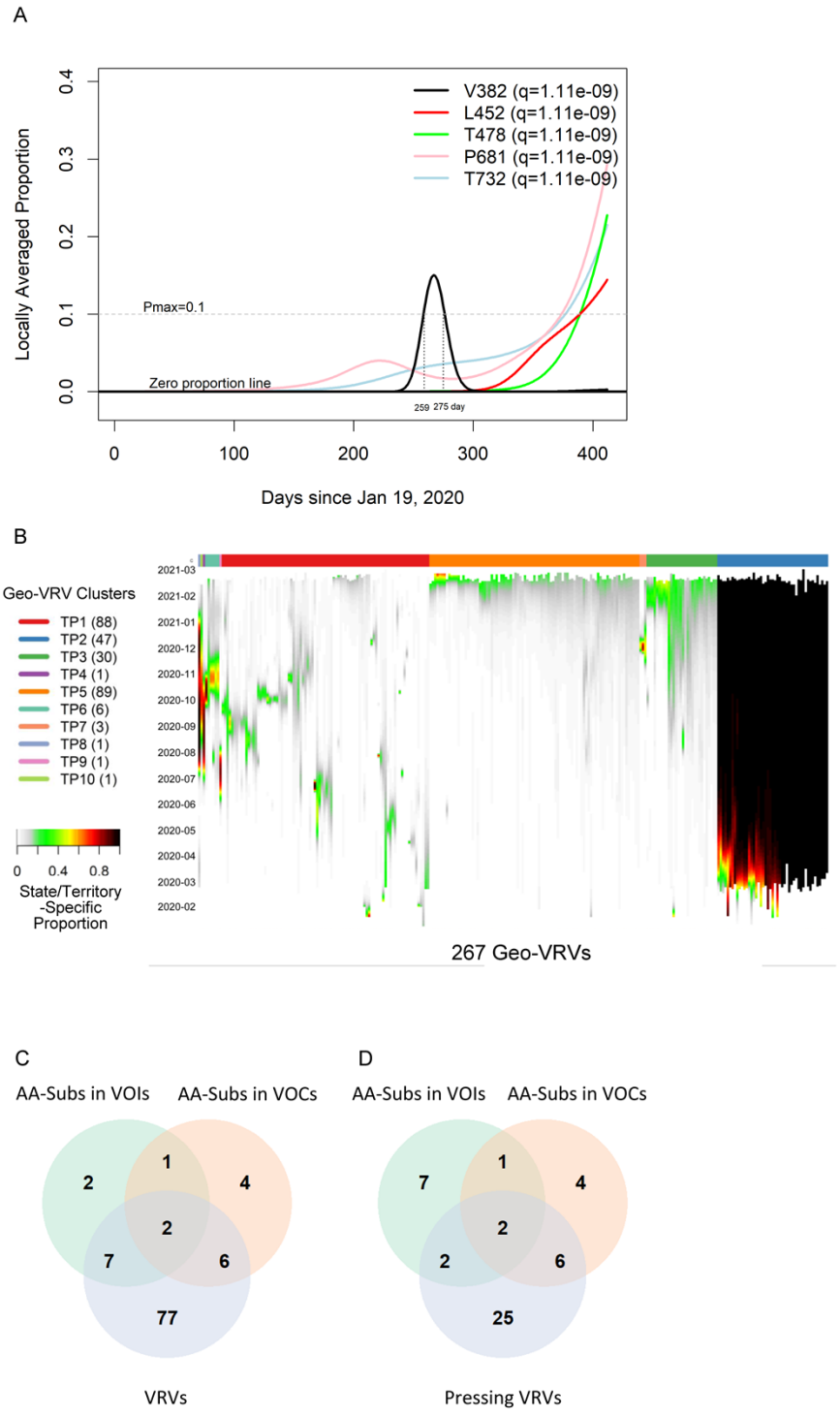


Fig. 1. Viral Residue Variant (VRV) spatiotemporal patterns in the United States. (A) Locally averaged proportions over time of five VRVs (V382, L452, T478, P681 and T732), modeled using sequences from California. The horizontal gray dotted line denotes

the Pmax cutoff of 10%. V382 exceeded the Pmax cutoff of 10% on day 259, and dropped below the Pmax cutoff of 10% on day 275 (marked by the vertical gray lines). (**B**) Heatmap of the 267 identified geo-VRVs, with color designating the state/territory-specific VRV proportion at the sampling time as designated on the left-hand vertical axis. Geo-VRVs with similar temporal dynamics are grouped into 10 clusters (TP1 through TP10), as designated by the color bar at the top of the heatmap. (**C, D**) Venn diagrams showing the relationships between AA-subs in VOIs, AA-subs in VOCs, and (**C**) VRVs or (**D**) pressing VRVs. AA-subs, amino acid positions that have been shown to harbor substitutions within US-circulating variants; VOCs, variants of concern; VOIs, variants of interest.

We refer to the combination of a VRV and a state/territory in which it was identified as a "geo-VRV". A total of 267 geo-VRVs, consisting of combinations of 90 VRVs identified among the 52 state/territory classifications, were identified (Table S3). Fifty-eight VRVs were only observed in one state/territory, whereas 32 were observed in two or more (Table S4).

Unsupervised learning was next applied to organize the 267 geo-VRVs into 10 clusters (TP1 through TP10) (Fig. 1B, Table S3). The cluster most strikingly different from the others was "TP2", which was composed of 47 geo-VRVs, each of which contained the D614 VRV at a maximum relative abundance of 100%, showing the early dominance of the D614 VRV in these states/territories. Clusters TP3, and TP5 include geo-VRVs of potential concern, since they include VRVs that appear to have emerged within the last few months in their specific states/territories. In contrast, most VRVs in the remaining clusters tended to expand and contract within relatively short times in a given state/territory, making such VRVs likely less important

from a public health perspective. We termed these 35 VRVs that were uniquely identified in Clusters TP2, TP3, and TP5 "pressing VRVs".

Comparison with AA Positions where Substitutions Have Been Identified Within US-Circulating VOIs and VOCs

We next compared the 90 VRVs and the 35 pressing VRVs with the 12 and 13 AA positions that have been shown to harbor substitutions (AA-subs) within US-circulating VOIs and VOCs, respectively (*2*). The 90 VRVs included 9 and 8 AA-subs in VOIs and VOCs, respectively; the 35 pressing VRVs included 4 and 8 AA-subs in VOIs and VOCs, respectively (Fig. 1C), even though all VOI/VOC sequences were excluded from the current analysis. Notably, 25 of the VRVs that have not been previously identified as an AA-sub in a VOI or VOC appear to have emerging trajectories, demonstrating the potential of the SLS method to identify novel Spike AA positions that may warrant further investigation/observation.

Five VOI/VOC AA-subs (Y144, F888, V1176, H69, K417) were not identified as a VRV. Fig. S3 shows the state/territory-specific relative abundances over time for states/territories where substitutions were identified at these 5 positions (albeit without meeting the statistical significance criteria for identification as a VRV). Our data suggest that, individually, these AA positions may be of less interest in US.

Timely Detection of Emerging VRVs

Timely detection of potentially fast-emerging VRVs, and conversely, identification of VRVs likely not of concern, are both important for informing public health guidelines and for influencing research priorities. Given the importance of timely detection, we use the first time when a Pmax of a VRV exceeds 10% as the first reportable time. For each out of the set of AA-subs within VOIs/VOCs that were also identified as VRVs, Table 1 compares within each

state/territory the time of detecting an emerging VRV as calculated by the SLS method vs. the first appearance of the AA-sub in the scientific literature. The SLS identified emerging VRVs in an average of 207 days, vs 299 days (average of reported values in literature). E484, an AA-sub in the B.1.1.7, P.1, and B.1.351 variants, is an exception as it was not detectable in the US until day 370, when it was first detected as a VRV in Rhode Island.

VRV-Haplotypes

SARS-CoV-2 is a single-stranded ("haploid") RNA virus. The presence of multiple VRVs found in a patient form a VRV-haplotype. The accumulation of multiple VRVs on a single RNA strand could affect protein function more than a single VRV. To identify VRV-haplotypes, we performed unsupervised learning of selected VRVs and cases through a two-way hierarchical cluster analysis state/territory-by-state/territory. As shown in Fig. S4, some VRV-haplotypes are shared across states/territories, but most are not. Fig. 2, for example, shows the results of the unsupervised case and VRV clustering for Washington state. The heatmap shows that multiple VRVs tend to aggregate among subsets of cases, inspection of which can reveal VRV-haplotypes as follows: The case cluster "PG8", which includes 14 cases, has VRVs from the "RG4" and "RG5" clusters, which include the VRVs (S13-W152-L452-V483-N501-D614-A684) (See Table S5).

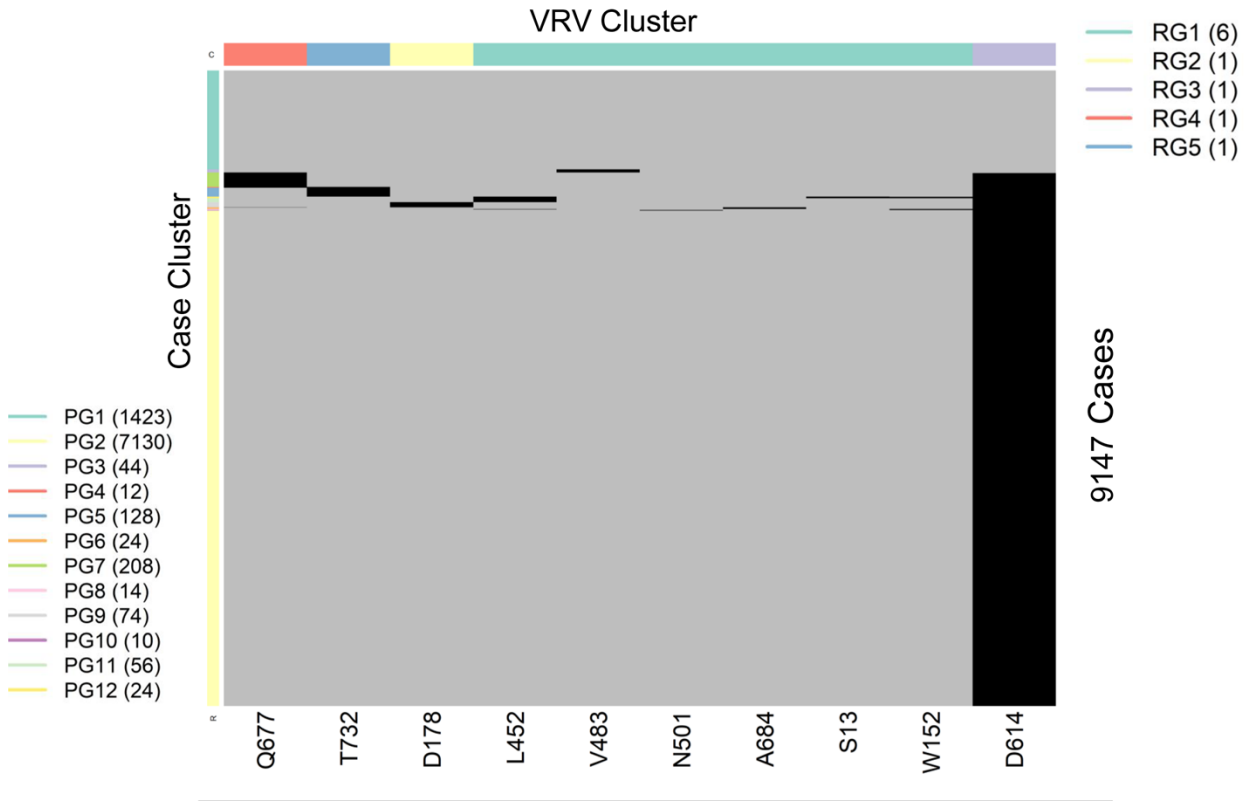


Fig. 2. Heatmap showing the presence of 10 selected VRVs among 9147 cases in Washington state. Unsupervised learning was used to organize the 10 VRVs into 5 residue groups (RG1 through RG5) and to organize the 9877 cases into 12 patient groups (PG1 through PG12).

Collectively, these four case clusters were combined to identify the VRV-haplotype W1 (Table 2), found in 104 cases in Washington. Similarly, the case cluster "PG4" (12 cases) had three VRVs (D614, Q677, T732) from the "RG3", "RG4", and "RG5" clusters. In total, six VRV-haplotypes (W1 through W6) were identified in Washington, while the "W6" cluster (7130 cases) carried only a single VRV, D614 (Table 2). Comparison across VRV-haplotypes

suggested that W6 evolved to W3, W4, and W5 via the acquisition of an additional mutation at T732, Q677, and D178, respectively. Similarly, both W3 and W4 could have evolved to W2 via the acquisition of an additional mutation at Q677 or T732, respectively.

VRV-haplotype blocks are identified from unsupervised learning. Within each block, there can be multiple VRV-haplotypes that consist of polymorphic residues; individual VRVs may take either the reference residue or a substitution. For example, VRV-haplotype W1 had 10 haplotypes (Table 2), where the number after the hyphen indicates the number of substitutions. For example, the haplotype "ICRVNGA" has four substitutions, and was observed twenty times in Washington.

Table 2 also displays the VRV-haplotypes observed in New York (N1 through N7). The most frequent block, N2, has seven VRVs and 16 unique haplotypes. Block N1 only differs from Block N2 via the acquisition of the P681 VRV, and thus the two blocks are closely connected. Similarly, Block N4, which probably gave rise to Block N3, has 14 unique haplotypes, including "GSRGNH" (six substitutions), which was observed 455 times. Lastly, N5 probably arose from N6 via N7, and has the "PGHI" haplotype (observed 367 times). We next used unsupervised learning to construct haplotypes in Washington and New York of the 35 pressing VRVs (Table S6).

Naming VRV Haplotypes via PANGO Lineages

As all sequences corresponding to VOI/VOC were excluded, the strains with detected VRVs are not currently undergoing special monitoring or characterization. We were thus interested in naming identified VRV-haplotypes and the PANGO lineages assigned by GISAID. To this end, we selected VRV-haplotype blocks including 4 or more pressing VRV mutations, resulting in 8 VRV-haplotype blocks. Table 3 cross-tabulates these VRV-haplotypes by their assigned

lineages. Of particular interest, viruses with the haplotype "KGHA" of T478-D614-P681-T732 were observed 2132 times, and 2029 of them were assigned to the strain B.1.1.222. It is natural to name the haplotype T478K-D614G-P681H-T732A as a B.1.1.222. Another noteworthy strain is B.1.234, which corresponds to "SVGHF" and "SVGHS" of G142-E180-D614-Q677-S940 with exceptionally high frequencies (353 and 262). The remaining VRV-haplotypes mostly correspond to B.1. Fourteen other strains were found in more than 10 occurrences and may also be of potential interest.

Impact of VRV Haplotypes on Viral Structure

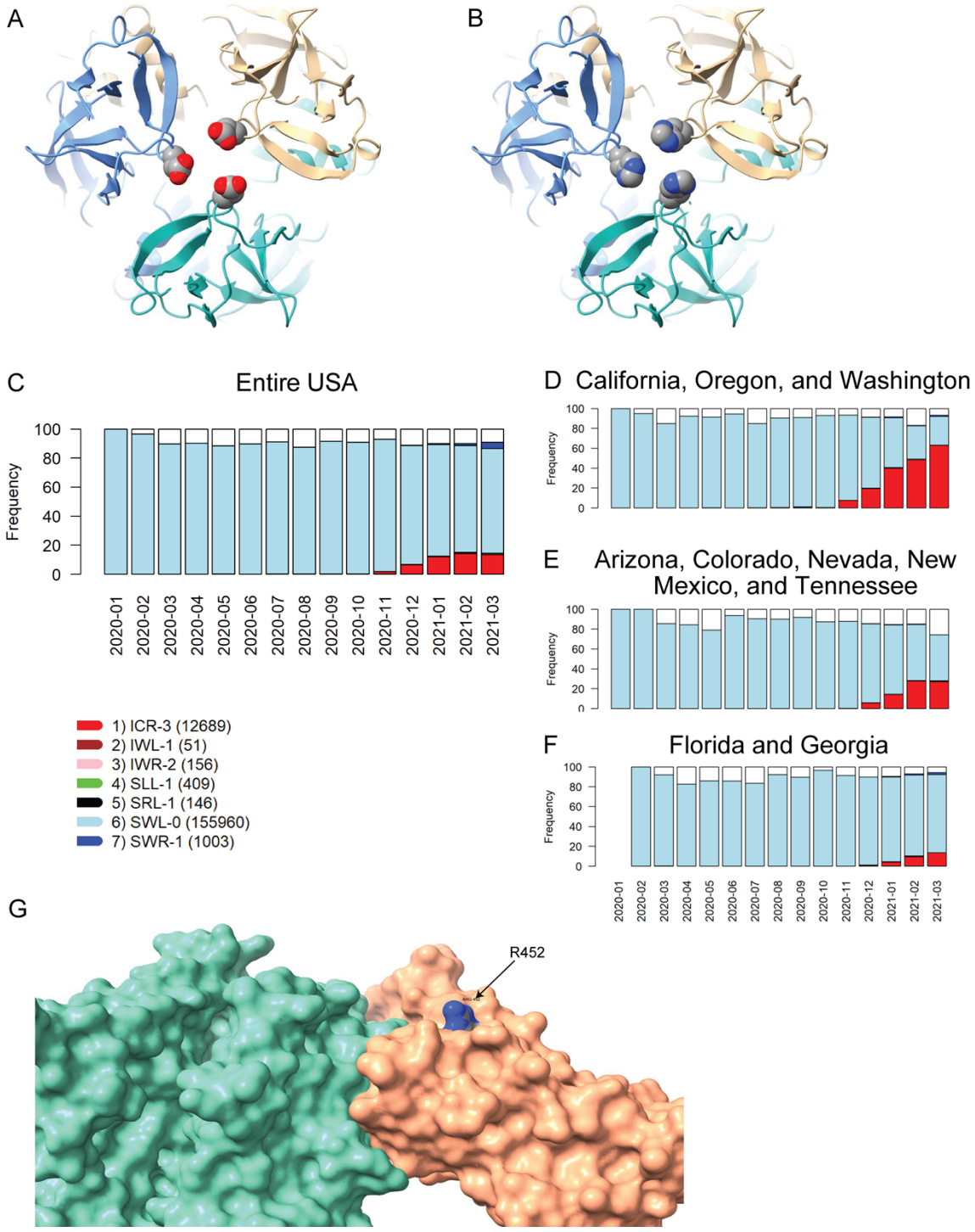
The SLS method includes homology modeling of Spike mutations, to predict possible consequences on Spike structure/function and to guide laboratory research. Inspection of the temporal dynamics of the VRV-haplotypes may be useful for identifying VRVs of interest. We performed homology modeling on two potentially interesting VRV-haplotypes, W1 (N501-A570-D614-P681-T716-S982-D1118, from the UK variant cluster B.1.1.7) and W2 [S13-W152-L452-D614, from the US variant cluster (B.1.94; B.1.427; B.1.429)].

The D614G mutation observed in the W1 haplotype has been associated with increased infectivity/transmissibility (*23-25*). Cryo-electron microscopy structures have been reported recently (*26, 27*) that reveal the structural consequences of this mutation and provide a plausible mechanistic explanation for the increased infectivity of D614G-carrying variants. The D614 VRV has predominated in all US cases for which sequence information is available in the TP2 cluster (Fig. 1B, Table S2). The N501Y mutation (present in the B.1.1.7 variant) is located in the receptor-binding domain (RBD) and has been reported to enhance binding affinity to the angiotensin-converting enzyme-2 (*10, 28*). N501Y has also been shown to reduce susceptibility

to some nAbs, although the B.1.1.7 variant appears to remain susceptible to some extent to natural infection-acquired and vaccine-induced nAbs (*10*).

Of the five remaining VRVs in the W1 haplotype, A570, T716, and S982 seem relatively benign in that mutations at these positions are already decreasing in certain states/territories (this trend is also true to some extent for N501Y). While this observation may simply reflect inadequate sequencing efforts in recent months, it may also indicate that mutations at these positions do not confer any fitness advantage to the virus.

The two remaining VRVs in the W1 haplotype, P681 and D1118, are more intriguing. Mutations at these two sites, particularly at P681, appear to persist in multiple states/territories. The P681H mutation occurs in the S1/S2 cleavage segment of the Spike protein, which is typically not resolved in cryo-electron microscopy or x-ray diffraction experiments. Thus, we cannot speculate on potential structural consequences of this mutation. However, the continued presence of this mutation in many states and its location in the Spike protein S1/S2 cleavage segment suggest that it may warrant further investigation. We are not aware of any reports that D1118H impacts transmissibility or morbidity, but the location of this mutation in the Spike protein trimer assembly (Fig. 3A, B) suggests it could impact trimer assembly structure/stability/dynamics.



246

247

248

249

Fig. 3. Homology modeling of Spike mutations and haplotypic polymorphisms over time of the S13-W152-L452 VRV-haplotype. (**A, B**) Modeled structure of the Spike protein trimer with (A) D1118 or (B) H1118 (homology-modelled using PDB entry 7KRS as the template structure). Spike protein monomers are displayed in blue, salmon, and aquamarine; aspartic acid and histidine residues are rendered as CPK images. (**C – F**) Frequencies over time for seven commonly observed haplotypic polymorphisms of the S13-W152-L452 VRV-haplotype, out of its polymorphisms in the US. Only haplotypic polymorphisms with at least 50 observations are included. Nomenclature is as follows: The first three letters designate the amino acids present at positions 13, 152, and 452, respectively; the number after the hyphen designates the number of amino acids at these three positions that do not match their reference strain equivalents. Numbers of sequences harboring each S13-W152-L452 haplotypic polymorphism (across the entire USA) are shown in parentheses. Frequencies of seven common S13-W152-L452 VRV-haplotypic polymorphisms (**C**) in the entire US; (**D**) in California, Oregon, and Washington combined; (**E)** in Arizona, Colorado, Nevada, New Mexico, and Tennessee combined; and (**F**) in Florida and Georgia combined. (**G**) Homology-modeled complex of the receptor-binding domain of the Spike protein (salmon), harboring the L452R mutation, bound to the angiotensin-converting enzyme 2 (ACE2) receptor (aquamarine). Within the R452 residue, nitrogen atoms are shown in blue and carbon atoms are shown in grey.

The US variants also carry the D614G mutation. The VRV-haplotype S13I-W152C-L452R (ICR-3) appeared in Fall 2020 and is rapidly becoming dominant in states on the West

Coast, as well as appearing in selected Southwestern and Southeastern states (Fig. 3C-3F). The S13I and W152C mutations, which are situated in the N-terminal domain (NTD) of the Spike protein, have been implicated in escape from NTD-targeting monoclonal antibodies (*29*). The L452R mutation is situated in the RBD; homology modelling of the RBD-ACE2 complex shows that while R452 does not directly contact ACE2, the guanidinium side chain of R452 is surface-exposed and thus could potentially impact nAb binding (Fig. 3G). The L452R mutation was recently shown to reduce binding affinity to some RBD-targeting monoclonal antibodies, as well as to reduce susceptibility to nAbs (*29*). Thus, structural modeling of mutations in the S13-W152-L452 VRV-haplotype yields results consistent with the temporal dynamics of this VRV-haplotype.

DISCUSSION

The continuous evolution of SARS-CoV-2 has already impacted public health guidelines and research priorities, with the potential of even more clinically consequential variants still to emerge. Here we leveraged a public data resource and described a statistical learning strategy for analyzing large, complex SARS-CoV-2 sequence datasets while incorporating temporal and spatial information. We provide detailed information on the emergence and persistence (or disappearance) of specific mutations in US states/territories, helping identify mutations that may warrant further observation/investigation. Our approach can be applied to other pathogens for which sufficient genomic surveillance data are available, generating important, statistically rigorous, and visually interpretable information for the biomedical research community, clinicians and public health officials. Our approach can also provide insight on the evolution of mutants and linkage with known viral strains.

By applying the SLS method to 167,893 US sequences not classified as any VOI/VOC, we identified 77 novel individual VRVs, including 25 pressing VRVs that appear to have emerged in the US. Among these pressing VRVs, the haplotype (T478-D614-P681-T732) links with the strain B.1.1.222 and (G142-E180-D614-Q677-S940) with the strain B.1.234, both of which do not correspond to any current VOI/VOC. Also of note, if the SLS method is applied to all US sequences, all circulating VOI/VOC are identified (results not shown).

As part of the assessment of immune correlates of protection, many randomized, placebo-controlled COVID-19 vaccine efficacy trials measure Spike protein sequences from symptomatic COVID-19 endpoint cases, and sometimes also from SARS-CoV-2 asymptomatic infections. Sieve analysis of these viral sequences can be conducted to assess whether and how vaccine efficacy depends on Spike protein sequence features, including differential vaccine efficacy across the levels of VRVs and of VRV-haplotypes (*30*). The graphical tools proposed here for spatiotemporal tracking of VRVs and VRV-haplotypes can be useful for sieve analysis, first by helping define and communicate the set of VRVs and VRV-haplotypes of study endpoint cases that have sufficient variability to be able to assess whether vaccine efficacy depends on the feature. For example, given that most vaccines use the Wuhan strain as the vaccine-insert, VRVs that meet our Pmax > 0.10 criterion would readily have the level of variability required for sieve analysis, whereas VRVs with Pmax < 0.02 would likely not. Secondly, including assignment to vaccine or placebo as a factor in the unsupervised clustering graphics applied to the vaccine efficacy trial sequence data sets may help communicate results of sieve analysis. Third, many of the vaccine efficacy trials have been offering the vaccine to placebo recipients, such that the placebo arm is lost and long term follow-up occurs only in individuals originally vaccinated or newly (deferred) vaccinated (*31*). The graphical tools may be applied to track study participant

vaccine breakthrough virus VRVs and VRV haplotypes over time, and to similarly track VRVs and VRV haplotypes in GISAID data bases of unvaccinated persons matched by geography and time, and a comparison of these two tracking results may aid sieve analysis during the long term follow-up period of the vaccine efficacy trials.

Evidence is mounting that neutralizing antibodies acquired by natural infection (*32, 33*) or through vaccination (*34, 35*) are a correlate of protection against COVID-19. Therefore, it will be critical to assess whether and how VRVs and/or VRV-haplotypes in the infecting strains impact neutralizing antibody titers attained by natural infection (*36*), as well as whether and how they impact neutralization sensitivity to vaccine-induced neutralizing antibodies (*12*) and/or monoclonal antibodies (*37*). One possibility is that the graphical tools used here could annotate VRVs and VRV-haplotypes according to impact on neutralization. Moreover, a subset of sieve analyses is designed to restrict to VRVs and VRV-haplotypes that are known to impact neutralization response to the given vaccine under study, to improve power and to contribute to understanding neutralizing antibody-based correlates of protection. Applications of pinpointing VRVs or VRV-haplotypes that impact vaccine efficacy, and to quantify their impact, include informing models for predicting vaccine efficacy against circulating virus populations, and to aid optimization of vaccine strain selection.

A limitation of our approach is that it is constrained by intrinsic sampling limitations, since all sequences were collected and contributed by laboratories without consistent sampling protocols. Hence, despite the large size of our dataset, the analyzed sequences were not nationally representative. Further, it is important to interpret our results in terms of VRV proportions among reported sequence data, rather than incidences or prevalence of VRVs, in the absence of reliably estimated denominators. To overcome this limitation, public health agencies

need to consider a uniformly developed surveillance protocol, to sequence COVID-19 cases from well-defined populations.

MATERIALS AND METHODS

Spike AA Sequences

Spike AA sequences (genome position: 21563-25384) from 189,727 COVID-19 cases in the US and selected US territories, along with their associated metadata, were retrieved from GISAID (*38*) (https://www.gisaid.org/) on March 23, 2021. Geographic origin (one of the 50 US states, Washington DC, Puerto Rico, or the Virgin Islands) was available for 189,284 of the sequences. For 443 of the cases, no US state/territory origin information was available. To ensure adequate sample size, Spike sequences from North Dakota, South Dakota, and the Virgin Islands were combined with these 443 sequences, forming an "Other States" category (728 sequences). Among them, 21,391 sequences were classified as a VOI or VOC (Table S1). These sequences were excluded, leaving 167,893 sequences for the analysis (see Table S2 for monthly case numbers by state/territory).

Sequence Alignment and Transformation to VRV Indicators

Spike protein sequences were aligned to the Wuhan reference sequence (*39*) using MAFFT (*40*), yielding a complete "rectangular residue sequence matrix". Sequences with at least one AA mutation (compared to the reference) were identified, enabling transformation of the residue sequence matrix to a matrix of binary VRV (mutant) indicators. Monomorphic residues led to columns of zeros and were eliminated from further analysis. We use VRV in this work to refer to a single AA position that harbors a substitution. We reserve the term "variant" in this work for identified VOIs and VOCs.

Statistical Learning Strategy (SLS)

Modeling VRV Temporal Dynamics

To model non-linear temporal dynamics, a generalized additive model (GAM) was used to regress the VRV indicator over sample collection time through a non-parametric regression model. Further details are given in the Supplementary Materials.

Visual Representation of Temporal Dynamics

Within-state/territory: Temporal dynamics of <8 VRVs within a given state/territory were visualized with a line plot. For visualizing temporal dynamics of ≥ 8 VRVs within a given state/territory, unsupervised learning was applied, grouping VRVs with similar temporal patterns. Results were visualized with a heatmap.

Spatially integrated: To visualize spatiotemporal VRV dynamics, all state-specific temporal dynamics were integrated and unsupervised learning (one-way hierarchical clustering with the Euclidean distance with weights in favor of recent temporal trajectories and the "ward.D2" agglomeration method) (*41*) was applied.

Missing Residue Imputation

Imputation of missing amino acid information is described in the Supplementary Material.

VRV-Haplotypes

A viral strain harboring multiple VRVs is referred to as a "VRV-haplotype". To identify VRV-haplotypes, unsupervised learning was used to organize both cases and VRVs through a two-way hierarchical analysis (*41*). Further information is given in the Supplementary Material.

Homology Modeling of Selected Haplotype Mutants

After identifying specific Spike protein mutants of interest from VRVs and related VRV-haplotypes, standard homology modeling methods were applied to generate 3D models. Further information is given in the Supplementary Material.

Supplementary Material

Table S5. VRV-haplotypes identified within each state/territory, along with state/territory-specific frequencies. The "positivity" column indicates the proportion of mutations in each haplotype block.

Table S6. Identified haplotypes of pressing VRVs in Washington and New York: frequencies, numbers of VRVs and haplotypic polymorphisms (frequency) in each state.

References and Notes

1. S. Duffy, Why are RNA virus mutation rates so damn high? *PLoS Biol* **16**, e3000003 (2018).

2. US Centers for Disease Control and Prevention, SARS-CoV-2 Variant Classifications and Definitions. https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-surveillance/variant-info.html Last updated 27 Apr, 2021. Access date 27 Apr, 2021.

3. X. Deng, M. A. Garcia-Knight, M. M. Khalid, V. Servellita, C. Wang, M. K. Morris, A. Sotomayor-Gonzalez, D. R. Glasner, K. R. Reyes, A. S. Gliwa, N. P. Reddy, C. Sanchez San Martin, S. Federman, J. Cheng, J. Balcerek, J. Taylor, J. A. Streithorst, S. Miller, G. R. Kumar, B. Sreekumar, P. Y. Chen, U. Schulze-Gahmen, T. Y. Taha, J. Hayashi, C. R. Simoneau, S. McMahon, P. V. Lidsky, Y. Xiao, P. Hemarajata, N. M. Green, A. Espinosa, C. Kath, M. Haw, J. Bell, J. K. Hacker, C. Hanson, D. A. Wadford, C. Anaya, D. Ferguson, L. F. Lareau, P. A. Frankino, H. Shivram, S. K. Wyman, M. Ott, R. Andino, C. Y. Chiu, Transmission, infectivity, and antibody neutralization of an emerging SARS-CoV-2 variant in California carrying a L452R spike protein mutation. *medRxiv*, (2021).

4. H. Tegally, E. Wilkinson, M. Giovanetti, A. Iranzadeh, V. Fonseca, J. Giandhari, D. Doolabh, S. Pillay, E. J. San, N. Msomi, K. Mlisana, A. von Gottberg, S. Walaza, M.

Allam, A. Ismail, T. Mohale, A. J. Glass, S. Engelbrecht, G. Van Zyl, W. Preiser, F. Petruccione, A. Sigal, D. Hardie, G. Marais, M. Hsiao, S. Korsman, M.-A. Davies, L. Tyers, I. Mudau, D. York, C. Maslo, D. Goedhals, S. Abrahams, O. Laguda-Akingba, A. Alisoltani-Dehkordi, A. Godzik, C. K. Wibmer, B. T. Sewell, J. Lourenço, L. C. J. Alcantara, S. L. K. Pond, S. Weaver, D. Martin, R. J. Lessells, J. N. Bhiman, C. Williamson, T. de Oliveira, Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. https://doi.org/10.1101/2020.12.21.20248640 *medRxiv*, 2020.2012.2021.20248640 (2020).

5. C. M. Voloch, R. da Silva Francisco, Jr., L. G. P. de Almeida, C. C. Cardoso, O. J. Brustolini, A. L. Gerber, A. P. C. Guimaraes, D. Mariani, R. M. da Costa, O. C. Ferreira, Jr., L. W. A. C. C. Covid19-Ufrj Workgroup, T. S. Frauches, C. M. B. de Mello, I. C. Leitao, R. M. Galliez, D. S. Faffe, T. Castineiras, A. Tanuri, A. T. R. de Vasconcelos, Genomic characterization of a novel SARS-CoV-2 lineage from Rio de Janeiro, Brazil. *J Virol*, (2021).

6. E. Volz, S. Mishra, M. Chand, J. C. Barrett, R. Johnson, L. Geidelberg, W. R. Hinsley, D. J. Laydon, G. Dabrera, Á. O'Toole, R. Amato, M. Ragonnet-Cronin, I. Harrison, B. Jackson, C. V. Ariani, O. Boyd, N. J. Loman, J. T. McCrone, S. Gonçalves, D. Jorgensen, R. Myers, V. Hill, D. K. Jackson, K. Gaythorpe, N. Groves, J. Sillitoe, D. P. Kwiatkowski, S. Flaxman, O. Ratmann, S. Bhatt, S. Hopkins, A. Gandy, A. Rambaut, N. M. Ferguson, Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data. 2020.12.30.20249034; doi:

https://doi.org/10.1101/2020.12.30.20249034 *medRxiv*, 2020.2012.2030.20249034 (2021).

7. E. Volz, S. Mishra, M. Chand, J. C. Barrett, R. Johnson, L. Geidelberg, W. R. Hinsley, D. J. Laydon, G. Dabrera, A. O'Toole, R. Amato, M. Ragonnet-Cronin, I. Harrison, B. Jackson, C. V. Ariani, O. Boyd, N. J. Loman, J. T. McCrone, S. Goncalves, D. Jorgensen, R. Myers, V. Hill, D. K. Jackson, K. Gaythorpe, N. Groves, J. Sillitoe, D. P. Kwiatkowski, C.-G. U. consortium, S. Flaxman, O. Ratmann, S. Bhatt, S. Hopkins, A. Gandy, A. Rambaut, N. M. Ferguson, Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature*, (2021).

8. P. Wang, M. S. Nair, L. Liu, S. Iketani, Y. Luo, Y. Guo, M. Wang, J. Yu, B. Zhang, P. D. Kwong, B. S. Graham, J. R. Mascola, J. Y. Chang, M. T. Yin, M. Sobieszczyk, C. A. Kyratsous, L. Shapiro, Z. Sheng, Y. Huang, D. D. Ho, Antibody Resistance of SARS-CoV-2 Variants B.1.351 and B.1.1.7. https://doi.org/10.1101/2021.01.25.428137 Access date 7 Mar, 2021. *bioRxiv*, 2021.2001.2025.428137 (2021).

9. D. A. Collier, A. De Marco, I. A. T. M. Ferreira, B. Meng, R. Datir, A. C. Walls, S. A. Kemp S, J. Bassi, D. Pinto, C. S. Fregni, S. Bianchi, M. A. Tortorici, J. Bowen, K. Culap, S. Jaconi, E. Cameroni, G. Snell, M. S. Pizzuto, A. F. Pellanda, C. Garzoni, A. Riva, A. Elmer, N. Kingston, B. Graves, L. E. McCoy, K. G. Smith, J. R. Bradley, J. James Thaventhiran, L. Lourdes Ceron-Gutierrez, G. Barcenas-Morales, H. W. Virgin, A. Lanzavecchia, L. Piccoli, R. Doffinger, M. Wills, D. Veesler, D. Corti, R. K. Gupta, SARS-CoV-2 B.1.1.7 escape from mRNA vaccine-elicited neutralizing antibodies. 2021.01.19.21249840; doi: https://doi.org/10.1101/2021.01.19.21249840 *medRxiv*, 2021.2001.2019.21249840 (2021).

480 10. P. Supasa, D. Zhou, W. Dejnirattisai, C. Liu, A. J. Mentzer, H. M. Ginn, Y. Zhao, H. M.
481 E. Duyvesteyn, R. Nutalai, A. Tuekprakhon, B. Wang, G. C. Paesen, J. Slon-Campos, C.
482 Lopez-Camacho, B. Hallis, N. Coombes, K. R. Bewley, S. Charlton, T. S. Walter, E.
483 Barnes, S. J. Dunachie, D. Skelly, S. F. Lumley, N. Baker, I. Shaik, H. E. Humphries, K.
484 Godwin, N. Gent, A. Sienkiewicz, C. Dold, R. Levin, T. Dong, A. J. Pollard, J. C.
485 Knight, P. Klenerman, D. Crook, T. Lambe, E. Clutterbuck, S. Bibi, A. Flaxman, M.
486 Bittaye, S. Belij-Rammerstorfer, S. Gilbert, D. R. Hall, M. A. Williams, N. G. Paterson,
487 W. James, M. W. Carroll, E. E. Fry, J. Mongkolsapaya, J. Ren, D. I. Stuart, G. R.
488 Screaton, Reduced neutralization of SARS-CoV-2 B.1.1.7 variant by convalescent and
489 vaccine sera. *Cell* **184**, 2201-2211 e2207 (2021).
490 11. S. A. Madhi, V. Baillie, C. L. Cutland, M. Voysey, A. L. Koen, L. Fairlie, S. D.
491 Padayachee, K. Dheda, S. L. Barnabas, Q. E. Bhorat, C. Briner, G. Kwatra, K. Ahmed, P.
492 Aley, S. Bhikha, J. N. Bhiman, A. E. Bhorat, J. du Plessis, A. Esmail, M. Groenewald, E.
493 Horne, S. H. Hwa, A. Jose, T. Lambe, M. Laubscher, M. Malahleha, M. Masenya, M.
494 Masilela, S. McKenzie, K. Molapo, A. Moultrie, S. Oelofse, F. Patel, S. Pillay, S. Rhead,
495 H. Rodel, L. Rossouw, C. Taoushanis, H. Tegally, A. Thombrayil, S. van Eck, C. K.
496 Wibmer, N. M. Durham, E. J. Kelly, T. L. Villafana, S. Gilbert, A. J. Pollard, T. de
497 Oliveira, P. L. Moore, A. Sigal, A. Izu, N.-S. G. W.-V. C. Group, Efficacy of the
498 ChAdOx1 nCoV-19 Covid-19 Vaccine against the B.1.351 Variant. *N Engl J Med*,
499 (2021).
500 12. W. F. Garcia-Beltran, E. C. Lam, K. St Denis, A. D. Nitido, Z. H. Garcia, B. M. Hauser,
501 J. Feldman, M. N. Pavlovic, D. J. Gregory, M. C. Poznansky, A. Sigal, A. G. Schmidt, A.

J. Iafrate, V. Naranbhai, A. B. Balazs, Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. *Cell*, (2021).

13. R. Rubin, COVID-19 Vaccines vs Variants-Determining How Much Immunity Is Enough. *Jama-J Am Med Assoc*, (2021).

14. D. A. Kennedy, A. F. Read, Monitor for COVID-19 vaccine resistance evolution during clinical trials. *PLoS Biol* **18**, e3001000 (2020).

15. D. M. Altmann, R. J. Boyton, R. Beale, Immunity to SARS-CoV-2 variants of concern. *Science* **371**, 1103-1104 (2021).

16. D. P. Maison, L. L. Ching, C. M. Shikuma, V. R. Nerurkar, Genetic Characteristics and Phylogeny of 969-bp S Gene Sequence of SARS-CoV-2 from Hawaii Reveals the Worldwide Emerging P681H Mutation. *bioRxiv*, (2021).

17. A. Rambaut, E. C. Holmes, A. O'Toole, V. Hill, J. T. McCrone, C. Ruis, L. du Plessis, O. G. Pybus, A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* **5**, 1403-1407 (2020).

18. J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, R. A. Neher, Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121-4123 (2018).

19. T. Koyama, D. Platt, L. Parida, Variant analysis of SARS-CoV-2 genomes. *Bull World Health Organ* **98**, 495-504 (2020).

20. E. C. Rouchka, J. H. Chariker, D. Chung, Variant analysis of 1,040 SARS-CoV-2 genomes. *PLoS One* **15**, e0241535 (2020).

21. K. M. Bindayna, S. Crinion, Variant analysis of SARS-CoV-2 genomes in the Middle East. *Microb Pathog* **153**, 104741 (2021).

22. D. M. Studdert, M. A. Hall, M. M. Mello, Partitioning the Curve - Interstate Travel Restrictions During the Covid-19 Pandemic. *N Engl J Med* **383**, e83 (2020).

23. P. Arora, S. Pohlmann, M. Hoffmann, Mutation D614G increases SARS-CoV-2 transmission. *Signal Transduct Target Ther* **6**, 101 (2021).

24. B. Korber, W. M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, N. Hengartner, E. E. Giorgi, T. Bhattacharya, B. Foley, K. M. Hastie, M. D. Parker, D. G. Partridge, C. M. Evans, T. M. Freeman, T. I. de Silva, C. McDanal, L. G. Perez, H. L. Tang, A. Moon-Walker, S. P. Whelan, C. C. LaBranche, E. O. Saphire, D. C. Montefiori, S. C.-G. Grp, Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812-+ (2020).

25. L. Z. Zhang, C. B. Jackson, H. H. Mou, A. Ojha, H. Y. Peng, B. D. Quinlan, E. S. Rangarajan, A. D. Pan, A. Vanderheiden, M. S. Suthar, W. H. Li, T. Izard, C. Rader, M. Farzan, H. Choe, SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nature Communications* **11**, (2020).

26. Z. Ke, J. Oton, K. Qu, M. Cortese, V. Zila, L. McKeane, T. Nakane, J. Zivanov, C. J. Neufeldt, B. Cerikan, J. M. Lu, J. Peukes, X. Xiong, H. G. Krausslich, S. H. W. Scheres, R. Bartenschlager, J. A. G. Briggs, Structures and distributions of SARS-CoV-2 spike proteins on intact virions. *Nature* **588**, 498-502 (2020).

27. J. Zhang, Y. Cai, T. Xiao, J. Lu, H. Peng, S. M. Sterling, R. M. Walsh, Jr., S. Rits-Volloch, H. Zhu, A. N. Woosley, W. Yang, P. Sliz, B. Chen, Structural impact on SARS-CoV-2 spike protein by D614G substitution. *Science*, (2021).

546 28. H. Liu, Q. Zhang, P. Wei, Z. Chen, K. Aviszus, J. Yang, W. Downing, C. Jiang, B.
547 Liang, L. Reynoso, G. P. Downey, S. K. Frankel, J. Kappler, P. Marrack, G. Zhang, The
548 basis of a more contagious 501Y.V1 variant of SARS-CoV-2. *Cell Res*, (2021).
549 29. M. McCallum, J. Bassi, A. Marco, A. Chen, A. C. Walls, J. D. Iulio, M. A. Tortorici, M.
550 J. Navarro, C. Silacci-Fregni, C. Saliba, M. Agostini, D. Pinto, K. Culap, S. Bianchi, S.
551 Jaconi, E. Cameroni, J. E. Bowen, S. W. Tilles, M. S. Pizzuto, S. B. Guastalla, G. Bona,
552 A. F. Pellanda, C. Garzoni, W. C. Van Voorhis, L. E. Rosen, G. Snell, A. Telenti, H. W.
553 Virgin, L. Piccoli, D. Corti, D. Veesler, SARS-CoV-2 immune evasion by variant
554 B.1.427/B.1.429. *bioRxiv*, (2021).
555 30. M. Rolland, P. B. Gilbert, Sieve analysis to understand how SARS-CoV-2 diversity can
556 impact vaccine protection. *PLoS Pathogens (in press)*, (2021).
557 31. D. Follmann, J. Fintzi, M. P. Fay, H. E. Janes, L. R. Baden, H. M. El Sahly, T. R.
558 Fleming, D. V. Mehrotra, L. N. Carpp, M. Juraska, D. Benkeser, D. Donnell, Y. Fong, S.
559 Han, I. Hirsch, Y. Huang, Y. Huang, O. Hyrien, A. Luedtke, M. Carone, M. Nason, A.
560 Vandebosch, H. Zhou, I. Cho, E. Gabriel, J. G. Kublin, M. S. Cohen, L. Corey, P. B.
561 Gilbert, K. M. Neuzil, A Deferred-Vaccination Design to Assess Durability of COVID-
562 19 Vaccine Effect After the Placebo Group Is Vaccinated. *Ann Intern Med*, (2021).
563 32. A. Addetia, K. H. D. Crawford, A. Dingens, H. Zhu, P. Roychoudhury, M. L. Huang, K.
564 R. Jerome, J. D. Bloom, A. L. Greninger, Neutralizing Antibodies Correlate with
565 Protection from SARS-CoV-2 in Humans during a Fishery Vessel Outbreak with a High
566 Attack Rate. *J Clin Microbiol* **58**, (2020).
567 33. A. G. Letizia, Y. Ge, S. Vangeti, C. Goforth, D. L. Weir, N. A. Kuzmina, C. A. Balinsky,
568 H. W. Chen, D. Ewing, A. Soares-Schanoski, M. C. George, W. D. Graham, F. Jones, P.

Bharaj, R. A. Lizewski, S. E. Lizewski, J. Marayag, N. Marjanovic, C. M. Miller, S. Mofsowitz, V. D. Nair, E. Nunez, D. M. Parent, C. K. Porter, E. Santa Ana, M. Schilling, D. Stadlbauer, V. A. Sugiharto, M. Termini, P. Sun, R. P. Tracy, F. Krammer, A. Bukreyev, I. Ramos, S. C. Sealfon, SARS-CoV-2 seropositivity and subsequent infection risk in healthy young adults: a prospective cohort study. *Lancet Respir Med*, (2021).

34. K. A. Earle, D. M. Ambrosino, A. Fiore-Gartland, D. Goldblatt, P. B. Gilbert, G. R. Siber, P. Dull, S. A. Plotkin, Evidence for antibody as a protective correlate for COVID-19 vaccines. https://doi.org/10.1101/2021.03.17.20200246 Posted 20 Mar, 2021. Access date 28 Apr, 2021. *medRxiv*, 2021.2003.2017.20200246 (2021).

35. K. McMahan, J. Yu, N. B. Mercado, C. Loos, L. H. Tostanoski, A. Chandrashekar, J. Liu, L. Peter, C. Atyeo, A. Zhu, E. A. Bondzie, G. Dagotto, M. S. Gebre, C. Jacob-Dolan, Z. Li, F. Nampanya, S. Patel, L. Pessaint, A. Van Ry, K. Blade, J. Yalley-Ogunro, M. Cabus, R. Brown, A. Cook, E. Teow, H. Andersen, M. G. Lewis, D. A. Lauffenburger, G. Alter, D. H. Barouch, Correlates of protection against SARS-CoV-2 in rhesus macaques. *Nature* **590**, 630-634 (2021).

36. D. Zhou, W. Dejnirattisai, P. Supasa, C. Liu, A. J. Mentzer, H. M. Ginn, Y. Zhao, H. M. E. Duyvesteyn, A. Tuekprakhon, R. Nutalai, B. Wang, G. C. Paesen, C. Lopez-Camacho, J. Slon-Campos, B. Hallis, N. Coombes, K. Bewley, S. Charlton, T. S. Walter, D. Skelly, S. F. Lumley, C. Dold, R. Levin, T. Dong, A. J. Pollard, J. C. Knight, D. Crook, T. Lambe, E. Clutterbuck, S. Bibi, A. Flaxman, M. Bittaye, S. Belij-Rammerstorfer, S. Gilbert, W. James, M. W. Carroll, P. Klenerman, E. Barnes, S. J. Dunachie, E. E. Fry, J. Mongkolsapaya, J. Ren, D. I. Stuart, G. R. Screaton, Evidence of escape of SARS-CoV-2

variant B.1.351 from natural and vaccine-induced sera. *Cell* **184**, 2348-2361 e2346 (2021).

37. R. E. Chen, X. Zhang, J. B. Case, E. S. Winkler, Y. Liu, L. A. VanBlargan, J. Liu, J. M. Errico, X. Xie, N. Suryadevara, P. Gilchuk, S. J. Zost, S. Tahan, L. Droit, J. S. Turner, W. Kim, A. J. Schmitz, M. Thapa, D. Wang, A. C. M. Boon, R. M. Presti, J. A. O'Halloran, A. H. J. Kim, P. Deepak, D. Pinto, D. H. Fremont, J. E. Crowe, Jr., D. Corti, H. W. Virgin, A. H. Ellebedy, P. Y. Shi, M. S. Diamond, Resistance of SARS-CoV-2 variants to neutralization by monoclonal and serum-derived polyclonal antibodies. *Nat Med* **27**, 717-726 (2021).

38. Y. L. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data - from vision to reality. *Eurosurveillance* **22**, 2-4 (2017).

39. C. Wang, Z. Liu, Z. Chen, X. Huang, M. Xu, T. He, Z. Zhang, The establishment of reference sequence for SARS-CoV-2 and variation analysis. *J Med Virol* **92**, 667-674 (2020).

40. T. Nakamura, K. D. Yamada, K. Tomii, K. Katoh, Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* **34**, 2490-2492 (2018).

41. F. Murtagh, P. Legendre, Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *J Classif* **31**, 274-295 (2014).

42. T. J. Hastie, *Generalized additive models. New York: Chapman and Hall, 1990*. (Chapman and Hall, New York, 1990).

43. J. D. Storey, J. E. Taylor, D. Siegmund, Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J R Stat Soc B* **66**, 187-205 (2004).

614 44. S. N. Wood, N. Pya, B. Safken, Smoothing Parameter and Model Selection for General
615 Smooth Models. *J Am Stat Assoc* **111**, 1548-1563 (2016).
616 45. P. Scheet, M. Stephens, A fast and flexible statistical model for large-scale population
617 genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J*
618 *Hum Genet* **78**, 629-644 (2006).
619 46. T. D. Goddard, C. C. Huang, E. C. Meng, E. F. Pettersen, G. S. Couch, J. H. Morris, T. E.
620 Ferrin, UCSF ChimeraX: Meeting modern challenges in visualization and analysis.
621 *Protein Sci* **27**, 14-25 (2018).
622 47. M. V. Shapovalov, R. L. Dunbrack, Jr., A smoothed backbone-dependent rotamer library
623 for proteins derived from adaptive kernel density estimates and regressions. *Structure* **19**,
624 844-858 (2011).
625 48. J. Lan, J. Ge, J. Yu, S. Shan, H. Zhou, S. Fan, Q. Zhang, X. Shi, Q. Wang, L. Zhang, X.
626 Wang, Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2
627 receptor. *Nature* **581**, 215-220 (2020).
628

629 **Acknowledgments:**

630 **Funding:**

631 This study was supported by the National Institutes of Health/National Institute of Allergy and
632 Infectious Diseases (https://www.niaid.nih.gov/) through award UM1 AI068635 to PBG.
633 The funders had no role in study design, data collection and analysis, decision to publish,
634 or preparation of the manuscript.

635 **Author contributions:**

636 Conceptualization: LPZ, TL, PBG, TRH, JTS, LS, THP, DEG, KRJ

637 Methodology: LPZ

638 Investigation: LPZ, TRH, JTS, LS, THP, DEG, KRJ

639 Visualization: LPZ

640 Funding acquisition: LPZ, PBG

641 Formal analysis: LPZ, TL

642 Data curation: JTS

643 Supervision: LPZ, PBG

644 Writing – original draft: LPZ, TL, PBG, JTS, LNC

645 Writing – review & editing: LPZ, PBG, TRH, JTS, LS, THP, LNC, DEG, KRJ

646 **Competing interests:** The authors declare that they have no competing interests.

647 **Data and materials availability:** All sequence data analyzed here are publicly available at

648 GSIAD (https://www.gisaid.org/).

Table 1. For 15 amino acid positions shown to harbor a substitution in a VOI or VOC, times estimated by the SLS method when the corresponding VRV had a locally averaged proportion exceeding 10% (and, if applicable, subsequently decreased below 10%) based on a state/territory-specific model. The top two rows show the first reported date in the literature of a VOI or VOC harboring a substitution at the designated site vs the date of VRV detection at the same amino acid position by the SLS method (across all states/territories).

	L5	S13	V70	T95	W152	D253	L452	S477	E484	N501	A570	D614	Q677	P681	A701
Reporting Day*	**301**	**301**	**301**	**301**	**301**	**87**	**301**	**331**	**87**	**362**	**362**	**362**	**362**	**362**	**362**
Earliest SLS Detection Day Across All States	**11**	**159**	**329**	**149**	**381**	**98**	**381**	**405**	**371**	**206**	**404**	**10**	**20**	**11**	**176**
Alabama	63-186								404-			63-	253-305-		
Alaska												56-	323	383-	
Arizona												26-	357-400	353-	
Arkansas												56-	329-		
California		398-			402-		390-					45-		374-	
Colorado	286-											45-	286-	370-	
Connecticut										314-		43-	191-	378-	
DC									391-			47-		344-	
Delaware												52-	384-	288-	
Florida												33-	368-	389-	
Georgia												41-	345-	407-	
Hawaii	175-190											46-	374-	174-376	
Idaho												53-			
Illinois												24-	366-	380-	
India												48-	370-	383-	
Iowa												48-	273-388		
Kansas						260-291						47-		385-	
Kentucky												59-	389-397		
Louisiana												50-	307-	368-	
Maine									404-	371-		51-			
Maryland					407-		394-		390-			45-		230-	
Massachusetts										206-		10-	346-	298-	
Michigan				149-177						264-273		50-	361-		
Minnesota	186-294											46-	297-	387-	
Mississippi	144-215											42-	353-	392-	
Missouri												47-		364-384	
Montana												68-			
Nebraska												46-		387-	

Nevada		392-			396-		393-						37-	391-	394-	
New Hampshire										374-			41-	349-390	364-	
New Jersey							402-						44-		276-	
New Mexico													50-	291-	387-	233-252
New York	11-13						386-		414-				11-		11-	
North Carolina													44-		382-	
North Dakota													57-	328-363		
Ohio							405-						20-	20-	391-	
Oklahoma													54-	306-		
Oregon		397-					398-						45-			
Pennsylvania													44-	394-	317-	
Puerto Rico						185-252							49-		347-	
Rhode Island								405-	371-384	356-			40-	358-398	379-	
South Carolina													46-	405-	368-389	
Tennessee	50-141												50-	318-		
Texas													23-	360-	378-	
Utah		159-173				98-190							44-		358-	
Virginia			329-331						397-				47-	384-	359-	176-185
Washington		406-			411-		403-			410-			50-	373-		
Wisconsin							409-			265-271			12-	299-407	411-	
Wyoming		381-			381-		381-						51-		392-	
Other States									393-	404-	404-		34-		368-	375-381

657

658

659 **"Reporting Day" was set to the 15th day in each month in which the relevant publication appeared.**

660 **"SLS Detection Day" was set to the day at which the locally averaged proportion of the specific VRV exceeded 10% based**
661 **on temporality models fitted in each states/territory. If the locally averaged proportion of the VRV later declined below**
662 **10%, the second day is shown after a hyphen.**

663 **All numbers in the table express the number of days post-January 19, 2021.**

664 **VOI, variant of interest; VOC, variant of concern.**

Table 2. VRV-haplotypes identified in Washington and in New York: state-specific frequencies of cases, number of VRVs per VRV-haplotype, and haplotypic polymorphisms (state-specific frequencies). Unimputable residues are denoted with an "X".

ID	VRV-haplotype	Freq	L	Haplotypic polymorphisms (frequency)
Washington				
W1	S13-W152-L452-V483-N501-D614-A684	104	4	ICRVNGA-4(20)/IWRVNGA-3(4)/SCRVNGA-3(5)/SLLVNGA-2(5)/ SRLVNGA-2(4)/SWLVTGA-2(4)/SWLVYDA-1(1)/SWLVYGA-2(5)/ SWQVNGA-2(2)/SWRVNGA-2(54)
W2	D614-Q677-T732	12	3	GHS-3(11)/XXX-3(1)
W3	D614-T732	128	2	GA-2(126)/GI-2(2)
W4	D614-Q677	208	2	DH-1(9)/GH-2(110)/GP-2(89)
W5	D178-D614	74	2	GG-2(70)/NG-2(4)
W6	D614	7130	1	G-1(7125)/N-1(5)
New York				
N1	L5-L54-E132-Y453-T478-E484-D614-P681-T732	172	9	LLEYKEGHA-4(168)/LLEYKEGHT-3(4)
N2	L5-L54-E132-Y453-T478-E484-D614-T732	651	8	FLEYREGT-3(4)/FLEYTEDT-1(11)/FLEYTEGA-3(3)/FLEYTEGS-3(1)/ FLEYTEGT-2(266)/FLEYTKGA-4(1)/FLEYTKGT-3(44)/LLEYKEGT-2(3)/ LLEYTAGT-2(1)/LLEYTEGA-2(51)/LLEYTEGI-2(2)/LLEYTEGS-2(24)/ LLEYTKGS-3(2)/LLEYTKGT-2(171)/LLEYTQGT-2(8)/LLQYTEGT-2(59)
N3	D80-F157-L452-D614-P681-T859-D950	132	7	DFLGHID-3(108)/DFLGPID-2(18)/DFLGPNH-3(4)/DSLGPNH-4(2)
N4	D80-F157-L452-D614-T859-D950	637	6	DFQGND-3(4)/DFRGID-3(15)/DFRGNH-4(1)/DFRGTD-2(120)/ DFRNTD-2(2)/DSRGNH-5(3)/DSRGTD-3(2)/GFRGND-4(1)/ GFRGNH-5(1)/ GSLGNH-5(9)/GSRGND-5(10)/GSRGNH-6(455)/ GSRGNY-6(1)/ GSRGTD-4(13)
N5	S494-D614-P681-T716	514	4	PGHI-4(367)/PGHT-3(55)/PGPT-2(52)/SGHI-3(19)/SGHT-2(8)/ SGPI-2(13)
N6	D614-P681	1161	2	GH-2(1124)/GL-2(4)/GR-2(32)/GS-2(1)
N7	D614	10822	1	D-0(1)/G-1(10821)

672 **Table 3. VRV-haplotypes.** Cross-tabulation of individual VRV-haplotypes with GISAID-
673 assigned lineages in all 167,893 sequences, excluding lineages with fewer than 10 occurrences.
674 "Freq", corresponding haplotype frequencies; "Unknown", sequences not assigned to any
675 lineage.

Hap-Load	Freq	Unknown	A.2.4	B.1	B.1.1	B.1.1.1	B.1.1.171	B.1.1.222	B.1.1.29	B.1.1.304	B.1.1.317	B.1.152	B.1.165	B.1.166	B.1.2	B.1.215	B.1.234	B.1.256	B.1.324	B.1.350	B.1.354	B.1.360	B.1.399	B.1.94
1) D80-F157-L452-D614-T859-D950																								
DSRGNH-5	63			58								5												
GSLGNH-5	9			9																				
GSRGND-5	21			19																			1	
GSRGNH-6	539			522					1			2		3									5	
2) D80-S155-F157-L452-T859-D950																								
DRSRNH-5	39			39																				
GRSRND-5	3			3																				
GRSRNH-6	30			30																				
GSSRNH-5	509			492					1			2		3									5	
3) G142-E180-D614-Q677-S940																								
SEGHF-4	3														1		1			1				
SVGHF-5	353																353							
SVGHS-4	273	2		1													262			8				
4) S155-F157-L452-T859-D950																								
RSRND-4	3			3																				
RSRNH-5	69			69																				
SSRNH-4	533			511					1			7		3									5	
5) S13-W152-L452-D614																								
ICLG-3	43	1		36											3									
ICRG-4	795	51		557									1	4	10			14			10	34	2	72
IWRG-3	120	1		77											7						2			28
SCRG-3	30	4		16											4									
6) S494-D614-P681-T716																								
PGHI-4	521			467	1									1	1	20				3				
PGHT-3	194			100	8			3					31		2	3			29				1	
RGHI-4	3															3								
SGHI-3	38			19	3			1							4									
7) T478-D614-P681-T732																								
KGHA-4	2132	11		17	2		14	2029	18	1	12				2									
KGHS-4	6																							
KGHT-3	159			4	57		3		67	8												1		
KGPA-3	5			1				3	1															
TGHA-3	85			13				63	2	1	2				2									
8) F157-L452-D614-T859																								

FQGN-3	22				22																
FRGI-3	15	14	1																		
FRGN-3	5		5																		
SLGN-3	11		10											1							
SRGN-4	625		601					1			7		3								6
SRGT-3	37		33																		

Green shading: > 100 occurrences. Light green shading: >10 occurrences.